Statistiek : kansen en verwachtingen

Neeleman, N.

2003

document version
Publisher's PDF, also known as Version of record

prof.dr.ir. N. Neeleman

Statistiek

Kansen en verwachtingen

vrije Universiteit *amsterdam*

prof.dr.ir. N. Neeleman

Statistiek

Kansen en verwachtingen

Rede uitgesproken op 16 januari 2003 ter gelegenheid van zijn afscheid als hoogleraar Statistiek aan de faculteit der Psychologie en Pedagogiek van de Vrije Universiteit Amsterdam.

vrije Universiteit *amsterdam*

Mijnheer de rector magnificus,

Dames en heren,

De ontwikkelingen in het statistiekonderwijs in de afgelopen dertig jaar, in het bijzonder de ontwikkelingen rond de basiscursus statistiek, zijn het onderwerp van mijn rede. Deze cursus is daarom zo belangrijk omdat hierin de grondslagen van de schattings- en toetsingstheorie behandeld worden waarop in eventuele vervolgcursussen wordt voortgebouwd. Onderzoek heeft geleerd dat, juist over deze grondslagen, er nogal wat misverstanden leven. Niet alleen onder studenten maar ook onder onderzoekers. Het lijkt mij daarom een goede gelegenheid hier nader in te gaan op die factoren die van invloed zijn op het ontstaan van deze misverstanden. Daartoe zal ik de diverse factoren zoals de docenten, het curriculum, de literatuur, de studenten en de onderzoekers in het kort de revue laten passeren. Daarbij zal ik tevens ingaan op de vele misverstanden die zowel bij de studenten als onderzoekers leven. Eén en ander zal ik illustreren met voorbeelden van foutieve en misleidende formuleringen in leerboeken. Ik zou hierbij willen opmerken dat de kanttekeningen die ik maak niet alleen gelden voor de basiscursus aan de VU, maar voor tal van soortgelijke cursussen in binnen- en buitenland.
Als motto voor mijn rede zou ik de woorden van Henry StJohn willen kiezen: "De waarheid is nauw begrensd, maar de ruimte voor vergissingen is onbeperkt."

De docenten
In 1940 verscheen van de hand van de beroemde statisticus Harold Hotelling een artikel in het blad "The Annals of Mathematical Statistics" getiteld "The Teaching of Statistics". Hierin beschrijft hij de geschiedenis van het statistiekonderwijs aan een faculteit X vanaf het moment dat deze faculteit het besluit heeft genomen dat een cursus statistiek in het onderwijsprogramma moet worden opgenomen.
Ook toen waren de tijden niet gunstig voor de universiteiten want de faculteit beslist dat één en ander niet teveel mag kosten, vandaar dat men geen statisticus aantrekt maar de cursus toevertrouwt aan de brillante, net aan deze faculteit afgestudeerde Jones, die al blijk heeft gegeven van een zekere bekwaamheid in het omgaan met getallen, zoals blijkt uit zijn scriptie die handelt over de toename van het aantal vierkante meters gravel op het universiteitsterrein.
Jones is een serieus mens en bereidt zich terdege voor op het geven van deze nieuwe cursus. Neuzend in de bibliotheek ontdekt hij het toonaangevende tijdschrift Biometrika. Hij voelt zich wat ongemakkelijk met het wiskundig niveau van de artikelen daarin. Ook de tot dan toe verschenen boeken zijn hem te diepzinnig. Jones realiseert zich dat er een markt is voor een elementaire inleiding in de statistiek en dat zodra hij wat meer kennis heeft verzameld hij in deze leemte kan voorzien. Zo gezegd zo gedaan en al spoedig blijkt dat Jones, een beter econoom dan statisticus, de markt goed heeft aangevoeld en zijn boek wordt een groot succes. Hij maakt promotie en wordt hoogleraar (in Hyphenated Statistics). Zijn boek, dat vele malen wordt herdrukt, wordt een inspiratiebron voor andere Jones's die zich, wat later dan hij, voor dezelfde taak gesteld zien. De boeken van deze auteurs bevatten niet alleen de fouten van Jones maar de meeste auteurs voegen er nog enkele van eigen hand aan toe. Hetgeen, zoals Hotelling opmerkt, de studie wie van wie iets heeft overgeschreven aanzienlijk vereenvoudigd. Als het bovendien een fout van Jones betreft die fout is overgeschreven is dat natuurlijk totaal onschadelijk.
Hotelling hekelt de gewoonte om in deze boeken de bewijzen achterwege te laten met het argument dat de studenten te weinig wiskunde kennen. Volgens hem is het eerder zo dat

de docenten de bewijzen zelf niet kunnen volgen. Hij komt tot de conclusie dat om een goed statistiekdocent te zijn men, ongeacht de vooropleiding, een grondige kennis van het onderwerp moet hebben, inclusief de wiskundige achtergrond, om alle mogelijkheden en beperkingen van de onderwezen technieken te kunnen overzien. Als de docent, wat vaak het geval is, anderen van statistisch advies dient, is deze eis nog klemmender.
Samengevat komt de conclusie van Hotelling neer op de slogan: "Men moet geen methoden onderwijzen waarvan men niet begrijpt hoe ze werken".
In 1982, dus 40 jaar later, werd op de Ohio State University een conferentie gehouden over "The teaching of Statistics and Statistical Consulting". Op deze conferentie was Geisser, een vooraanstaand statisticus, één van de sprekers. Hij refereerde aan het door mij hiervoor aangehaalde artikel van Hotelling en kwam na bespreking van het statistiekonderwijs op dat tijdstip tot de conclusie dat er nog steeds veel klonen van Jones rondlopen. Het lijkt mij dat dit anno 2003 nog steeds het geval is.

Laat ik, ter illustratie, u hiervan een tweetal voorbeelden geven. Beide voorbeelden zijn ontleend aan de variantieanalyse. Een veel gebruikte techniek waarbij men aan de hand van een aantal steekproeven uit evenzovele verschillende populaties wil beslissen of de populaties identiek zijn of dat twee of meerdere populaties van elkaar verschillen.
Voorwaarde om deze techniek te mogen toepassen is dat de populaties normaal verdeeld zijn en dat de populatievarianties gelijk moeten zijn.
Nu is niets in het leven volmaakt en in de praktijk wordt aan deze eisen zelden voldaan. Belangrijk is dan ook welke afwijkingen van deze voorwaarden men mag toestaan voordat bij toepassing van deze techniek de uitkomsten ontoelaatbaar kunnen worden beïnvloed.
Nu is er in 1954 van de hand van Box een artikel verschenen waarin werd aangetoond dat, indien de populaties normaal verdeeld zijn en de grootste en kleinste populatievariantie niet meer dan een factor drie van elkaar verschillen, men rustig een variantieanalyse kan uitvoeren mits de steekproeven allemaal dezelfde omvang hebben.
In 1974 verscheen van de hand van Brown en Forsythe een artikel waarin werd aangetoond dat indien aan deze voorwaarden niet wordt voldaan de variantieanalyse een onbetrouwbare techniek is. Een conclusie die bevestigd werd in de artikelen van Rogan en Keselman (1977), Tomarken en Serlin (1986) en Wilcox, Charlin en Thompson (1986).
Toch vond ik, al of niet naar het artikel van Box verwijzend de volgende uitspraken in leerboeken die aan de opleidingen psychologie in gebruik zijn.

1. *Als de populaties symmetrisch verdeeld zijn en de grootste en kleinste variantie niet meer dan een factor vier à vijf van elkaar verschillen is deze techniek robuust mits men werkt met gelijke steekproefomvang. (Howell 1987)*

De grens wordt naar boven toe bijgesteld en de eis van normaalverdeelde populaties wordt verruimd.

2. *Zolang de grootste en de kleinste steekproefomvang niet meer dan een factor 1½ verschillen is het niet belangrijk of de varianties verschillen. (Hays 1988)*

Ook hier wordt noch over een grens noch over normaal verdeelde populaties gesproken.

3. *Als de varianties verschillen maar de steekproeven qua omvang gelijk zijn of slechts weinig verschillen is de variantieanalyse robuust. (Neter, Kutner, Nachtsheim en Wasserman 1996)*

Over een grens wordt zelfs niet meer gesproken, evenmin als over normaal verdeelde populaties

4. *Als de grootste en kleinste variantie niet meer dan een factor vier verschillen mag*

men zonder meer variantieanalyse toepassen. (Moore and McCabe 1999)
Dat, afgezien van het feit dat de grens naar boven wordt bijgesteld, in ieder geval de steekproeven gelijke omvang behoren te hebben en afkomstig behoren te zijn uit normaal verdeelde populaties wordt in het geheel niet vermeld.

Zoals u ziet wordt de door Box gegeven factor drie opgerekt tot vier, vijf of wordt er zelfs geen grens meer genoemd. Dit is des te opmerkelijker omdat Wilcox in een artikel in 1989 aantoonde dat in de psychologie en pedagogie in bepaalde vakgebieden de grootste en de kleinste steekproefvarianties vaak een factor 16 of meer van elkaar verschillen. In één publicatie zelfs werd een factor van 121 gevonden.
Tevens wordt er geheel aan voorbij gegaan dat de populaties normaalverdeeld dienen te zijn.

Een tweede voorbeeld is het dwingend voorschrijven van een tweestap beslissingsprocedure bij de variantieanalyse. In de eerste stap, wordt aan de hand van de zogenaamde F-toets beslist of men concludeert of de populatiegemiddelden verschillen. In de tweede stap wordt met behulp van een zogenaamde een multiple comparison toets nagegaan welke populatiegemiddelden verschillen.

1. *Na de F-toets komt posthoc vergelijking van populatiegemiddelden met behulp van de methode van Tukey of Scheffé. (Hays 1988)*
2. *De F-toets is een toets die voorafgaat aan een nadere analyse van de verschillen tussen de populatiegemiddelden. (Neter, Kutner, Nachtsheim en Wasserman 1996)*
3. *Men moet de F-toets toepassen voordat men een multiple comparison toets mag toepassen. (Moore and McCabe 1999)*

In 1973 verscheen een artikel van de hand van Olshen waarin betoogd werd dat de kans om populatiegemiddelden ten onrechte als verschillend aan te wijzen aanzienlijk hoger was als de Scheffé procedure vooraf werd gegaan door een significante F-toets dan indien men de betreffende techniek zonder voorafgaande F-toets toepaste.
In 1975 verscheen een simulatiestudie van Bernhardson waaruit bleek dat deze kans voor een vijftal andere multiple comparison procedures eveneens aanzienlijk hoger was indien deze procedures vooraf werden gegaan door een significante F-toets dan indien men de betreffende techniek zonder voorafgaande F-toets toepaste.

Het lijkt mij dat in 2003 de conclusie van Hotelling dat men een grondige kennis van de statistiek dient te bezitten, inclusief de wiskundige achtergrond, voordat men zich aan het geven van de basiscursus waagt, nog steeds geldig is.

Het curriculum en de literatuur
Daar verreweg de meeste leerboeken op het gebied van de statistiek voor sociale wetenschappen de stof van de basiscursus behandelen wil ik deze twee onderwerpen gezamenlijk behandelen.
Over het algemeen komen er in de vloed van inleidende boeken op het gebied van de statistiek voor de sociale wetenschappen nogal wat onjuistheden voor. Zo gaf Brewer in 1985 in het Journal of Educational Statistics een uitgebreide bespreking van de in dat jaar verschenen inleidende werken op het gebied van de statistiek voor sociale wetenschappen en de daarin voorkomende fouten.
Hij constateerde dat deze fouten betrekking hadden op hypothesentoetsen,

betrouwbaarheidsintervallen, de centrale limietstelling en de verdeling van het steekproefgemiddelde, dus praktisch de gehele stof.
Dat de studenten, naast de moeilijkheden die ze met de stof hebbem, vaak op het verkeerde been worden gezet door onduidelijke en/of onjuiste formuleringen in de leerboeken, moge duidelijk zijn.

Het lijkt mij hier de plaats in te gaan op een aantal specifieke moeilijkheden waarmee studenten te kampen hebben bij de basiscursus statistiek. Eén en ander zal ik illustreren met voorbeelden van de door Brewer gevonden fouten.
De basiscursus statistiek is opgebouwd uit drie delen te weten beschrijvende statistiek, kansrekening en inductieve statistiek. Bij het onderwijs wordt, bij de behandeling van de stof, meestal ook deze volgorde aangehouden. Qua moeilijkheidsgraad verschillen de onderdelen echter aanzienlijk van elkaar.
Voor verreweg de meeste studenten levert het onderdeel beschrijvende statistiek, waar het gaat om waarnemingsmateriaal op een nette wijze in tabellen samen te vatten en in grafieken weer te geven en om het berekenen van een aantal kentallen, zoals gemiddelde en variantie, weinig tot geen moeilijkheden op.
Het volgende onderdeel, de kansrekening, heeft een veel hogere moeilijkheidsgraad. Niet voor niets zegt Charles Sander Pierce, de Amerikaanse filosoof:"This branch of Mathematics (Probability) is the only one I believe,in which good writers frequently get results entirely erroneous." Zo berekende de grote wiskundige Leibniz de kans om in twee worpen met een dobbelstenen 11 te gooien verkeerd omdat hij de elementaire fout beging geen rekening te houden met het feit dat men een zes en een vijf niet alleen in de volgorde (6,5) maar ook (5,6) kan gooien.

De kansrekening levert het fundament waarop het laatste onderdeel, de inductieve statistiek, wordt voortgebouwd. Indien men niet voldoende tijd besteedt om de kansrekening grondig te behandelen dan zijn voor de studenten, moeilijkheden met het gedeelte inductieve statistiek onvermijdelijk.

Op een tweetal onderwerpen uit de kansrekening die voor goed begrip van de inductieve statistiek van belang zijn en die tot veel misverstanden aanleiding geven, wil ik hier kort ingaan.
Het gaat hier om voorwaardelijke kansen en gecombineerd de normale verdeling en de centrale limietstelling. Met name voor de klassieke schattings- en toetsingstheorie zijn deze onderwerpen van belang.
Bij de behandeling van het begrip voorwaardelijke kans zijn de drie meest voorkomende misvattingen bij de studenten de volgende:

1. *Men denkt bij voorwaardelijke kansen aan causaliteit.*
   Een voorbeeld ter toelichting.
   Stel men heeft een vaas met twee zwarte en twee witte ballen. Uit deze vaas neemt een proefleider na elkaar twee ballen. De volgende vragen worden nu aan de studenten gesteld. Wat is de kans dat de tweede bal die de proefleider pakt zwart is, als gegeven is dat de eerste bal die hij gepakt heeft wit was. Wat is de kans dat de eerste bal die de proefleider gepakt heeft zwart was als u weet dat de tweede bal die hij gepakt heeft wit was.

   De studenten hebben geen moeilijkheid met het antwoord op de eerste vraag $\frac{2}{3}$.

   Zij redeneren dat als de eerste trekking wit is, er nog twee zwarte en één witte bal over zijn. Bij het antwoord op de tweede vraag geeft de overgrote meerderheid aan

dat deze kans $\frac{1}{2}$ is onder de motivering dat er twee witte en twee zwarte ballen in de vaas zitten en het de eerste bal het niet kan schelen of de tweede bal wit of zwart is. Zij zien niet in dat, gezien vanuit het standpunt van de informatievoorziening, de situatie symmetrisch is en dat het goede antwoord $\frac{2}{3}$ is.

Men gaat uit niet van de gegeven voorwaarde maar van een daarvan afgeleide voorwaarde. Een voorbeeld

Er zitten drie fiches in een vaas één is aan beide zijden wit één is aan beide zijden rood en één is aan de ene zijde rood en aan de andere zijde wit. Men trekt een fiche uit de vaas en legt hem op tafel. De bovenzijde is rood. Wat is de kans dat de onderzijde ook rood is?

De meeste studenten zeggen dat deze kans $\frac{1}{2}$ is, uitgaande van de afgeleide voorwaarde dat het fiche met de twee witte zijden het niet kan zijn en dat beide andere fiches evenveel kans hebben om getrokken te worden. Men dient echter uit te gaan van alle fiches. In totaal zijn er dan zes zijden die boven kunnen liggen, drie daarvan zijn rood en drie daarvan zijn wit. Van de drie waarbij een rode zijde boven ligt zijn er twee waarbij ook een rode zijde onderligt en is er één waarbij een witte zijde onderligt. De correcte kans is dus $\frac{2}{3}$.

3. *De verwarring met de inverse kans.*

Als voorbeeld een opgave uit de basiscursus.

Stel dat het percentage hoogbegaafde kinderen is 0.3% en dat men beschikt over een instrument waarmee men kan meten of iemand hoogbegaafd is. Men heeft dit instrument uitgeprobeerd op een groot aantal hoogbegaafde kinderen en 99% van deze kinderen werden door dit instrument aangewezen als hoogbegaafd. Tevens heeft men dit instrument uitgeprobeerd op een groot aantal niet hoogbegaafde kinderen en 95% van die kinderen werden door dit instrument aangewezen als niet hoogbegaafd.

De voorwaardelijk kans dat iemand door het instrument wordt geïdentificeerd als hoogbegaafd, gegeven dat hij hoogbegaafd is, is dus gelijk aan 0.99.

Evenzo is de voorwaardelijk kans dat iemand door het instrument wordt geïdentificeerd als niet hoogbegaafd, gegeven dat hij niet hoogbegaafd is, gelijk aan 0.95.

In de praktijk is men echter geïnteresseerd in de inverse kans namenlijk de kans dat iemand hoogbegaafd is, gegeven dat het instrument hem als hoogbegaafd heeft aangewezen. Deze inverse kans is slechts 0.05623 en wijkt dus aanzienlijk af van de oorspronkelijke kans. Evenzo kan men berekenen dat de kans dat iemand niet hoogbegaafd is, gegeven dat het instrument hem als niet hoogbegaafd heeft aan gewezen, gelijk is aan 0.99997.

Het door elkaar halen van de oorspronkelijke voorwaardelijke kansen en de bijbehorende inverse kansen treedt bij studenten en niet alleen bij hen veelvuldig op.

Het tweede onderwerp dat een belangrijke rol speelt in de basiscursus statistiek is de normale verdeling. Zowel de klassieke toetsingstheorie als de schattingstheorie, punt- zowel als intervalschatting, zijn gebaseerd op de veronderstelling dat men beschikt over aselect getrokken steekproeven uit normaal verdeelde populaties. Daarnaast is de centrale limietstelling mede oorzaak van deze dominante rol van de normale verdeling. Hoewel het bewijs van deze stelling geen deel uitmaakt van de basiscursus, is het relatief

eenvoudig de inhoud van deze stelling te verduidelijken aan de hand van een computersimulatie of zelfs met een aantal eenvoudige berekeningen met een krijtje op een bord. In het kort komt de stelling hierop neer dat onder zeer algemene voorwaarden de verdeling van het steekproefgemiddelde van een aselecte steekproef uit een willekeurig verdeelde populatie goed benaderd kan worden door een normale verdeling met parameters het populatiegemiddelde en de populatievariantie gedeeld door de steekproefomvang, mits deze steekproefomvang groot genoeg is. Sommige versies in de tekstboeken bevatten het magische getal 25 of 30, daarbij stellend dat de benadering door de normale verdeling voldoet als de steekproefomvang de 25 of 30 passeert. Een veel voorkomend misverstand is dat de studenten denken dat als de steekproefomvang groter is dan 30 men zich geen zorgen behoeft te maken over de vorm van de populatieverdeling en men de klassieke toetsings- en schattingsrecepten, die gebaseerd zijn op normaal verdeelde populaties zonder meer mag toepassen. Sommigen gaan daarbij zelfs zover dat, indien de exacte verdeling bekend is, ze toch de normale benadering gebruiken. Als zelfs bij een correcte weergave van het centrale limiet theorema deze misverstanden optreden hoeveel temeer zal men fouten maken als dit theorema foutief is weergegeven. Brewer geeft o.a. de volgende voorbeelden:

1. *A distribution of means of samples of equal size, as described above, when taken from an infinite population, will form a normal population.*
2. *As the size of the samples increases, the t-distribution becomes more and more normal in shape and when samples are of infinite size, they are similar to the normal distribution.*
3. *In section 10.2 we said that from the basic result of the central limittheorem we learn that any variable that is resampled repeatedly and randomly tends to be distributed normally, the larger the size of the sample is.*
4. *We can generalize from this and say that the size of the standard error of any statistic is inversely proportional to the number of cases in the sample upon which the statistic was computed.*

Het gedeelte inductieve statistiek bevat van de drie onderdelen van de basiscursus de meest complexe redeneringen. Het loont om, bij de aanvang van dit gedeelte van de cursus, uitgebreid stil te staan bij de begrippen steekproefgrootheid en steekproefverdeling.

Het is voor veel studenten namelijk bijzonder moeilijk het steekproefgemiddelde te zien als een kansvariabele met een verdeling en de waarde van het steekproefgemiddelde als een realisatie van deze kansvariabele. Aan de andere kant dient de student er zich van bewust te zijn dat het populatiegemiddelde een constante is. Verwarring op dit punt leidt tot veel voorkomende fouten bij studenten. Bekend is de volgende uitspraak over een betrouwbaarheidsinterval: " De kans dat het populatiegemiddelde in dit interval ligt is 95%", waarbij een constante, het populatiegemiddelde, als een kansvariabele wordt behandeld.

Het is voor de studenten moeilijk zich te realiseren dat, bij intervalschatting, de kansen op het berekeningsmechanisme slaan en niet op een individuele realisatie.

Als voorbeelden van missers in de leerboeken geeft Brewer o.a.

1. *In our example, the 99% confidence interval is then 73 ± 3.87 or 69.13 to 76.87. When we assert that the unknown μ falls within the range of values, 99% of such assertions will be correct*
2. *When we use a 99% interval, we have the chance of only 1 to 100 of being wrong.*

De meeste misverstanden treden bij de studenten op bij het onderwerp hypothesentoetsen. Ook hier speelt de verwarring tussen steekproefgemiddelde en populatiegemiddelde een grote rol. Vaak stelt de student hypothesen op over steekproefgemiddelden en spreekt over een significant verschil tussen de gemiddelden van twee steekproeven.
Verder is het van belang de student er op te wijzen dat de onbetrouwbaarheid van een toets, de kans om de nulhypothese te verwerpen gegeven dat deze nulhypothese juist is, een voorwaardelijke kans is. Studenten worden hier vaak op het verkeerde been gezet omdat men in veel boeken deze onbetrouwbaarheid presenteert als de kans om de nulhypothese ten onrechte te verwerpen maar daarbij niet het voorwaardelijk karakter van deze kans benadrukt.
Daarnaast is een veel voorkomend misverstand de al eerder besproken verwisseling van deze kans voor zijn inverse kans. Men interpreteert dan de onbetrouwbaarheid als de kans dat de nulhypothese juist is gegeven dat de nulhypothese wordt verworpen. Deze inverse kans is zonder verdere veronderstellingen echter niet te berekenen maar kan aanzienlijk verschillen van de onbetrouwbaarheid. (Denk aan het voorbeeld van de hoogbegaafden).

1. *An α level of 0.05 implies that the probability of making a type 1 error bij chance alone is 5 in 100.*
2. *When we are rejecting a hypothesis at the 1 percent level, we are saying that the chances are 99 in 100 that it is false.*

De vraag van Brewer, "Zijn statistiekboeken voor de sociale wetenschappen bronnen van mythen en foutieve opvattingen?" kan dunkt mij in veel gevallen bevestigend beantwoord worden.

De inhoud van de basiscursus statistiek is verrassend lang ongewijzigd is gebleven. Verhoudingsgewijs recent zijn zaken als onderscheidingsvermogen en in het verlengde daarvan de bepaling van de optimale steekproefomvang in het curriculum opgenomen.
Een recente ontwikkeling in de literatuur is het opnemen van een gedeelte dataverwerking in het curriculum, dit gaat meestal ten koste van de het deel kansrekening. Als voorbeeld noem ik hier het boek "Early Succes in Statistics" waar de hele kansrekening is teruggebracht tot een halve pagina! Gemiddelde en variantie worden, na een summiere uitleg ingevoerd als toetsaanslagen. Men voert de data in en interpreteert volgens vaste richtlijnen de output.
Zoals een Amerikaanse collega onlangs zei: "Het is erg verleidelijk de statistiekcursus om te bouwen tot een cursus waarin de studenten een statistisch pakket leren gebruiken. Dit leidt echter zonder meer tot het "pluginski syndrome." Men voert domweg zonder nadenken data in de computer in en geeft een computercommando.
Het bezwaar tegen een dergelijke aanpak is dat men snel verouderende vaardigheden leert en geen blijvende kennis verwerft noch zelfstandig leert nadenken.

Indien men de aan de schattings- en toetsingstheorie ten grondslag liggende ideeën goed beheerst, is niets op tegen om de verschillende toetsen in de vorm van een recepten te presenteren. Is dat echter niet het geval dan leidt één en ander tot grote misverstanden. Als men geen ei kan bakken is het niet verstandig om te gaan koken.

De studenten
Tijdens een bijeenkomst van docenten statistiek over de stand van zaken over computerondersteund statistiekonderwijs zei bij zijn inleiding de heer Sijtsma:"Iedere statistiekdocent weet uit eigen ervaring dat zijn vak door vele studenten vooral wordt

gezien als een noodzakelijke en vooral lastige horde op weg naar de begeerde bul. Ook vragen de studenten zich wel eens af wat het nut is van een vak als statistiek voor bijvoorbeeld de psychologie. Wiskunde gaat toch niet over mensen, dus hoe zou statistiek iets zinnigs kunnen bijdragen aan het begrip van mentale processen? Naast deze bekende motivatieproblemen vinden veel studenten statistiek gewoon een moeilijk vak waarvoor zij harder moeten werken dan voor veel andere vakken. Daarmee lijkt statistiek zo ongeveer de status te hebben gekregen van wiskunde op de middelbare school: moeilijk en niet echt leuk, en bovendien valt er niet gemakkelijk in te zien wat je er nu precies aan hebt. Overigens dient gezegd te worden dat vele studenten wel plezier hebben in dit uitdagende vak."

Hetzelfde speelt in de USA waar een spreker tijdens een conferentie over de statistiek opmerkte. "Voorwaarde voor onze cursus is het behaald hebben van het high school diploma. Het vreemde verschijnsel doet zich nu voor dat geen van de studenten, zelfs de eerstejaars niet, de laatste vier jaren algebra heeft gehad. Tenminste daar lijkt het op. Onze studenten zijn zwak in rekenen en veel van de motivatieproblemen bij onze cursus worden veroorzaakt door het gevoel van de studenten dat zij in wiskundig opzicht tekort schieten".

Nu zijn de wiskunde-eisen die aan de basiscursus statistiek gesteld worden zeer beperkt. Zo bevat de cursus geen enkele afleiding of bewijs en wordt er qua wiskunde van de student niet meer verwacht dan het invullen van een formule en een zekere rekenvaardigheid. Dat het met deze rekenvaardigheid niet best gesteld is moge blijken uit het feit dat 80% van de studenten op de vraag van $\frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.99 \times 0.05}$ *stelde dat dit gelijk was aan* $\frac{1}{0.99 \times 0.05}$*, en dat onder de resterende 20% er nog enkele waren die nul als uitkomst opgaven. Dit is helemaal in lijn met het bericht in een landelijk ochtendblad dat een groot deel van de leerlingen van de PABO niet in staat zijn een voldoende te halen voor de testen voor taal en rekenen in de hoogste klassen van de lagere school.*

Het is niet de bedoeling met dit voorbeeld de studenten belachelijk te maken maar het gaat mij erom te laten zien op welk wiskundeniveau de studenten na het behalen van het VWO diploma met wiskunde A1 binnenkomen. Wat ze op het VWO gehad hebben heeft weinig met wiskunde van doen. Het is allemaal stuitend concreet. Elke som wordt teruggebracht een praktische ervaring van de leerling. In de gehele stof van A1 is geen bewijs te vinden. Alles wordt al of niet aan de hand van illustraties en voorbeelden aannemelijk gemaakt.

Er wordt geen inzicht bijgebracht, vergelijkingen oplossen, grafieken tekenen, nulpunten en uiterste waarden bepalen worden nu met de grafische rekenmachine opgelost. Je hoeft het niet meer te snappen om het te kunnen oplossen. De aankomende generatie studenten kan wel oplossingen produceren maar weet niet waarvan het de uitkomsten zijn. Wat nodig is om in de wiskunde vorderingen te maken is oefening. Wiskundige concepten zijn geconstrueerd uit oude concepten die op een nieuwe zinvolle wijze worden samengesteld. Maar deze oude concepten zijn weer samengesteld uit nog oudere concepten. Men kan geen calculus leren als men de algebraïsche vaardigheden niet onder de knie heeft en men kan deze weer niet leren als men geen rekenvaardigheid bezit. Wiskunde is rücksichtslos cumulatief. Het inoefenen van bepaalde vaardigheden wordt door de onderwijskundigen gezien als slecht voor het begrip van de zaak. Een gedeelte van de wiskunde beheersen geeft veel voldoening, maar het is een beloning voor hard en niet altijd even plezierig werk. Zoals de grote natuurkundige Feynman eens zei:"Zonder parate kennis kan je natuurkunde niet snappen." en hetzelfde geldt voor de

wiskunde.
Wie geen routine heeft in bepaalde zaken lijkt op een fietsenmaker die weigert om een fiets op te bouwen uit een frame, kant en klare wielen, en verlichtingsset enzovoort maar eerst deze componeneten zelf in elkaar zet.
Het zal duidelijk zijn dat de eerstejaars, die nu binnenkomen met wiskunde A1 een lager wiskundeniveau dan de studente "oude stijl" die wiskunde A in hun pakket hadden en deze hebben weer een lager niveau dan indertijd de studenten met wiskunde 1.
Sinds de jaren tachtig van de vorige eeuw daalt het wiskundeniveau van de eerstejaars en de door de universiteit aan hen gestelde toelatingseiseneisen dalen mee.
Was het nog zo dat in de jaren tachtig 80% van de studenten in één jaar slaagden voor hun statistiek, na invoering van wiskunde A slaagden er, ondanks aanpassing van het niveau en invoering van werkgroepen, nog maar 60% in één jaar. Met ingang van dit cursusjaar is de wiskundeniveau verder verlaagd tot wiskunde A1. Dit gevoegd bij de eis dat de cursus statistiek in één derde van de tijd die er in de tachtiger jaren voor werd uitgetrokken, moet zijn afgerond, leidt onvermijdelijk tot een verdere, mijns inziens onverantwoorde daling van het niveau van de basiscursus. Het is typisch een geval van de boerenwens minder kippen meer eieren. De geruststellende mededeling die ik las dat één en ander door een betere didactiek zal worden opgevangen is wishful thinking. De beste didactiek kan een dergelijke daling van wiskundeniveau en reductie in tijd niet opvangen.

De onderzoekers
Voor onderzoekers is de statistiek een van de belangrijkste werktuigen. Het is aannemelijk dat zij met het oog op onderzoek meer verdiept hebben in dit vak dan de gemiddelde student. Bovendien hebben zij door hun rol als van scriptiebegeleider en mentor van assistenten in opleiding een grote invloed op de ideeën van deze jeugdige onderzoekers. In het eerder geciteerde "Early Succes in Statistics" stellen de auteurs:"De staf bestaat uit wijze mensen waarnaar je altijd goed moet luisteren. Als ze bepaalde toetsen aanraden, waarvan je geleerd hebt dat deze toetsen met je gegevens onverenigbaar zijn, dan hebben ze daar vast goede redenen voor. Bovendien moeten ze straks je scriptie of je proefschrift beoordelen."

Bij bestudering van artikelen en proefschriften komen een drietal zaken naar voren.
Ten eerste
Als men het overgrote deel van hun publicaties beschouwd dan kan het haast niet anders dan dat ze de uitspraak van Hopkins en Glass in hun boek Basic statistics for the behavioral sciences: onderschrijven. Deze schrijvers stellen: "Het is een gelukkig toeval dat vele variabelen in allerlei disciplines verdelingen bezitten die goed door een normale verdeling benaderd kunnen worden. Anders gezegd: " God houdt van de normale verdeling"
De opvattingen van Geary zijn, mijns inziens, echter aanzienlijk realistischer. Geary stelt in zijn artikel Testing for normality dat: "tengevolge van het brillante werk van R.A.Fisher die aantoonde dat, als men normaliteit veronderstelde, men uit steekproeven van welke omvang dan ook conclusies van het grootste praktische belang kon trekken de vooringenomenheid ten gunste van de normale verdeling in volle hevigheid terugkeerde. Het belang van de daaraan ten grondslag liggende veronderstellingen werd bijna geheel vergeten. Hij vervolgt met:" Normaliteit veronderstellende steekproefgrootheden kunnen relatief niet-robuust zijn indien men te maken heeft met niet-normale verdelingen. Ondanks dat zijn de leerboeken en de researchliteratuur van de sociale wetenschappen

doordrongen van de veronderstelling van normaliteit." Tot slot merkt hij op:"normaliteit is een mythe er was nooit en er zal nooit zoiets zijn als een normale verdeling."

Deze opvatting wordt krachtig ondersteund door het onderzoek van Micceri (1989). Deze onderzoeker verzocht auteurs van recente goed bekend staande tijdschriften waarin grootschalig onderzoek werd gepubliceerd, hun gegevens aan hem ter beschikking te stellen teneinde na te gaan of de bij hun onderzoek gebruikte steekproeven afkomstig waren uit normaal verdeelde populaties. Op deze wijze verzamelde hij 440 steekproeven van een omvang van tussen de 200 en 10000 stuks. Van de vervaardigde histogrammen waren er 312 ééntoppig, 89 twee toppig en 39 had meer dan twee toppen. De auteur vergeleek deze steekproeven op een aantal maten met de numeriek bepaalde maten voor steekproeven van deze omvang uit normaal verdeelde populaties. Daarnaast toetste hij, met een kleine onbetrouwbaarheid, of deze steekproeven afkomstig waren uit normaal verdeelde populaties. Geen van de 440 beoordeelde steekproeven bleek afkomstig uit een normaal verdeelde populatie. Wat de auteur met zorg vervulde dat, hoewel het steeds ging om recent gepubliceerd onderzoek 75% van de auteurs, om de meest uiteenlopende redenen, niet in staat waren hun gegevens ter beschikking te stellen.
Men zou denken dat een dergelijk uitgebreid onderzoek toch zijn impact zou hebben op de gebruikte statistische technieken van onderzoekers. Niets is echter minder waar. Nog steeds wordt het overgrote deel van gepubliceerd onderzoek uitgegaan van normaal verdeelde populaties zonder dat de steekproeven daar ooit op gescreend zijn.

Ten tweede
Veel onderzoekers zijn, zoals Tversky en Kahneman het noemen, overtuigd van de wet van de kleine aantallen. Zij bedoelen daarmee te zeggen dat ze te veel vertrouwen stellen in de uitkomsten van te kleine steekproeven.
Om dit aan te tonen stelden Tversky en Kahnemman aan de deelnemers van een conferentie een zestal vragen, Als voorbeeld heb ik een van deze vragen geselecteerd.

> *Veronderstel dat u met behulp van een instrument aan 20 proefpersonen een meting heeft uitgevoerd. Gegeven is dat deze metingen normaal verdeeld zijn en dat de standaardafwijking van het instrument bekend is. U toetst tweezijdig ($\alpha = 0.05$) en vindt een significant resultaat ($z = 2.23$). U besluit het onderzoek te herhalen en meet met hetzelfde instrument 10 nieuwe proefpersonen. Deze keer toetst u eenzijdig en kiest de alternatieve hypothese in de richting van het eerder gevonden resultaat ($\alpha = 0.05$).*
> *Hoe groot schat u de kans dat u bij deze toets weer een significant resultaat vindt?*

Het overgrote deel van de congresbezoekers schatte deze kans in op zo'n 70 tot 80%.Het goede antwoord is 50%. Ook de antwoorden op de andere vragen gaven een soortgelijk antwoordpatroon te zien.
Ondersteuning van hun stelling kregen Tversky en Kahneman toen in 1969 het bekende boek van Cohen "Statistical Power Analysis for the Behavioral Sciences" verscheen.
Hierin ontwikkelde Cohen voor een aantal bekende veelgebruikte technieken een methode om, uitgaande van bepaalde veronderstellingen, een optimale steekproefomvang te berekenen. Niet te groot zodat onbeduidende verschillen worden ontdekt, wat het onderzoek onnodig duur maakt, maar ook niet te klein zodat de kans op ontdekking van interessante verschillen te klein wordt. Volgens het uitgebreide onderzoek van Cohen is bij verreweg de meeste onderzoeken de kans een effect van gemiddelde grootte te ontdekken ongeveer 50%. Ofwel als er een effect van gemiddelde grootte aanwezig is kan

men net zo goed met een geldstuk gooien en stellen als kruis bovenkomt is er een effect, als munt bovenkomt is er geen effect.
Het onderzoek van Cohen werd in 1989 door Sedlmeier en Gigerenzer herhaalt. De auteurs kwamen tot de conclusie dat ten opzichte van 1969 niets is veranderd.
Daarnaast concludeerden de auteurs dat studies waar voor de nulhypothese de researchhypothese werd gekozen, zelfs bij een gering aantal waarnemingen, geen vragen opriepen bij de referees en zonder meer geplaatst werden.
Dit roept de vraag op of men zelfs artikelen uit gerefereerde tijdschriften wel kan vertrouwen.
Ook hier is onderzoek naar gedaan. In 1999 liet Rossi, om zijn studenten ervan te overtuigen dat de door hem behandelde technieken ook echt door onderzoekers werden toegepast,. door studenten van een aantal artikelen waarvan voldoende gegevens voorhanden waren, de toetsingsgrootheden narekenen. Hij kwam tot de onthutsende conclusie dat bij 25 % van de in totaal 67 toetsen de afwijking tussen de gepubliceerde en de opnieuw berekende toetsingsgrootheid gevonden meer dan 20% bedroeg. Maar dit terzijde.

Ten derde
De laatste opmerking die ik zou willen maken is dat onderzoekers de neiging hebben om tot steeds complexere modellen hun toevlucht te nemen zoals LISREL, multilevelanalyse enz.. Deze modellen gaan er van uit dat de variabelen (multi)normaal verdeeld zijn. Over de robuustheid voor afwijkingen van normaliteit is bij deze modellen nog zeer weinig bekend.
Hierop is het volgende verhaal dat ik onlangs van een collega hoorde van toepassing
Op weg naar een conferentie ontmoeten vier onderzoekers en vier statistici elkaar in de trein.
Ze raken aan de praat en na een tijdje komt het gesprek op de prijs van de treinreis. "Nou", zeggen de statistici "dat valt nogal mee, wij reizen samen op één kaartje". Hoe doen jullie dat dan met de controle willen de onderzoekers weten. "O "zeggen de statistici "daar weten we wel raad op". Als enige tijd later de conducteur in de verte aankomt staan de statistici op en verdwijnen in het toilet en als de conducteur op de deur klopt schuiven ze hun enige kaartje onder de deur door.
Op de terugreis treffen de statistici en de onderzoekers elkaar weer in de trein. "We hebben van jullie geleerd hoor," zeggen de onderzoekers "wij reizen nu ook op één kaartje net als jullie". "Wij hebben nu helemaal geen kaartje" zeggen de statistici. "Hoe doen jullie dat dan bij de controle" vragen de onderzoekers. "O", zeggen de statistici "daar weten wij wel raad mee". Als in de verte de conducteur nadert staan de onderzoekers en de statistici op en verdwijnen in de tegenover elkaar liggende toiletten. Vlak voordat de conducteur arriveert klopt één van de statistici op de deur van het toilet van de onderzoekers waarop deze hun kaartje onder de deur door schuiven. De statisticus pakt het kaartje op en verdwijnt in zijn eigen toilet. De rest laat zich raden.
Moraal van dit verhaal: Pas geen methoden toe die je niet volledig begrijpt.

Statistiek, kansen en verwachtingen.
Uit het voorafgaande blijkt dat er vele factoren zijn, die de oorzaak kunnen zijn van de misverstanden die bij studenten en ook wel bij onderzoekers leven. Op een aantal zaken kan men invloed uitoefenen maar een zeer belangrijke oorzaak van het geringe kennisrendement van de basiscursus is het lage wiskundeniveau van de nu aankomende studenten. Hierover heeft de faculteit geen zeggenschap. De kans dat dit niveau op korte

termijn zal stijgen lijkt mij nihil. Wie de verwachting heeft dat, bij dit niveau onder de huidige nevenvoorwaarden een verantwoorde basiscursus gegeven kan worden, staat met beide benen stevig in de lucht.

Ik heb gezegd.

Geraadpleegde literatuur

Bernhardson, C. (1975). "Type I error rates when multiple comparison procedures follow a significant F test of ANOVA," Biometrics 50, 719-724.

Box, G.E.P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way model," Annals of Mathematical Statistics 25, 290-302.

Brewer J.K. (1985)." Behavioural statistics textbooks: source of myths and misconceptions," Journal of Educational Statistics 10, 252-268.

Brown, M.B. and Forsythe, A. (1974). "The small sample behavior of some statistics which test the equality of several means," Technometrics 16, 129-132.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, (2 nd ed.) Hillsdale N.J.: Erlbaum

Falk R. (1988). "Conditional probabilities: insights and difficulties." In: Davidson R, Swift J. (eds) Proceedings 2 nd International Conference on Teaching Statistics. University of British Colombia, Canada. 292-265.

Geary R.C. (1947). "Testing for normality", Biometrika 34, 209-242.

Geisser S. (1982)." Observations on graduate programs in statistics and related issues". In: Rustage J.S., Wolfe D.N. (eds) Teaching Statistics and Statistical Consulting. Academic Press. New York

Glickman L.V. (1990)." Lessons in counting from the history of probability," Teaching Statistics 12 15-17

Hawkins A., Jolliffe F., Glickman L. (1992). Teaching Statistical Concepts. Longman. London.

Hays W.L. (1988). Statistics. Holt, Rinehart and Winston. New York.

Hopkins K.D., Glass G.V. (1978). Basic Statistics for the Behavioral Sciences. Englewood Cliffs. N.J. Prentice Hall.

Hotelling H. (1940). " The teaching of statistics," Annals of Mathematical Statisics 11, 457-472.

Howell D.C. (1987). Statistical Methods for Psychology. Duxberry Press. Boston.

Maltby J., Day L (2002). Early Success in Statistics. Prentice Hall. London

Micceri, T. (1984). "The unicorn, the normal curve and other improbable creatures," Psychological Bulletin 105, 156-166.

Moore D.S., McCabe G.P. (1999). Introduction to the Practice of Statistics (3th ed.). W.H. Freeman and Company. New York.

Neter J., Kutner M.H., Nachtsheim C.J., Wasserman W. (1996). Applied Linear Statistical Models (4th ed.). Irwin. Chicago.

Olshen R.A. (1973)." The conditional level of the F-test," Journal of the American Statistical Associaton 68, 692-698.

Oskamp S. (1965)." Overconfidence in case-study judgments," Journal of Consulting Psychology 29, 261-265.

Rogan J.C.,Keselman H.J. (1977). "Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation," American Educational Research Journal 14, 493-498.

Sedlmeier P., Gigerenzer G. (1989)." Do studies of statistical power have an effect on the power of studies?" Psychological Bulletin 105, 309-316

Steiger J.H., Foulardi R.T. (1982)." Noncentrality Interval estimation and evaluation of statistical models," In: Rustage J.S., Wolfe D.N. (eds) Teaching Statistics and Statistical Consulting. Academic Press. New York.

Tversky A, Kahneman D. (1971)." Belief in the law of small numbers," Psychological Bulletin 76, 105-110

Tomarken A., Serlin R. (1986). "Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures," Psychological Bulletin 99, 90-99.

Wilcox R.R (1989). "Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models," Journal of Educational Statistics 14, 260-278.

Wilcox R.R, Charlin V., Thompson K.L. (1986). "New monte carlo results on the robustness of the ANOVA F, W, and F^* statistics," Communications in Statistics- Theory and Methods 15, 933-944.

Winkler R.L., Hays W.L. (1978). Statistics: Probability, Inference and Decision. Holt, Rinehart and Winston. New York.

derzoeken doorgeven veronderstellen bevestigen luistere

even veronderstellen bevestigen luisteren verwonderen waarnemen verwijzen vergelijken verbinde

n doorgeven veronderstellen bevestigen luisteren

eken doorgeven veronderstellen bevestigen luisteren verwonderen waarnemen verwijzen vergelijken verbinden

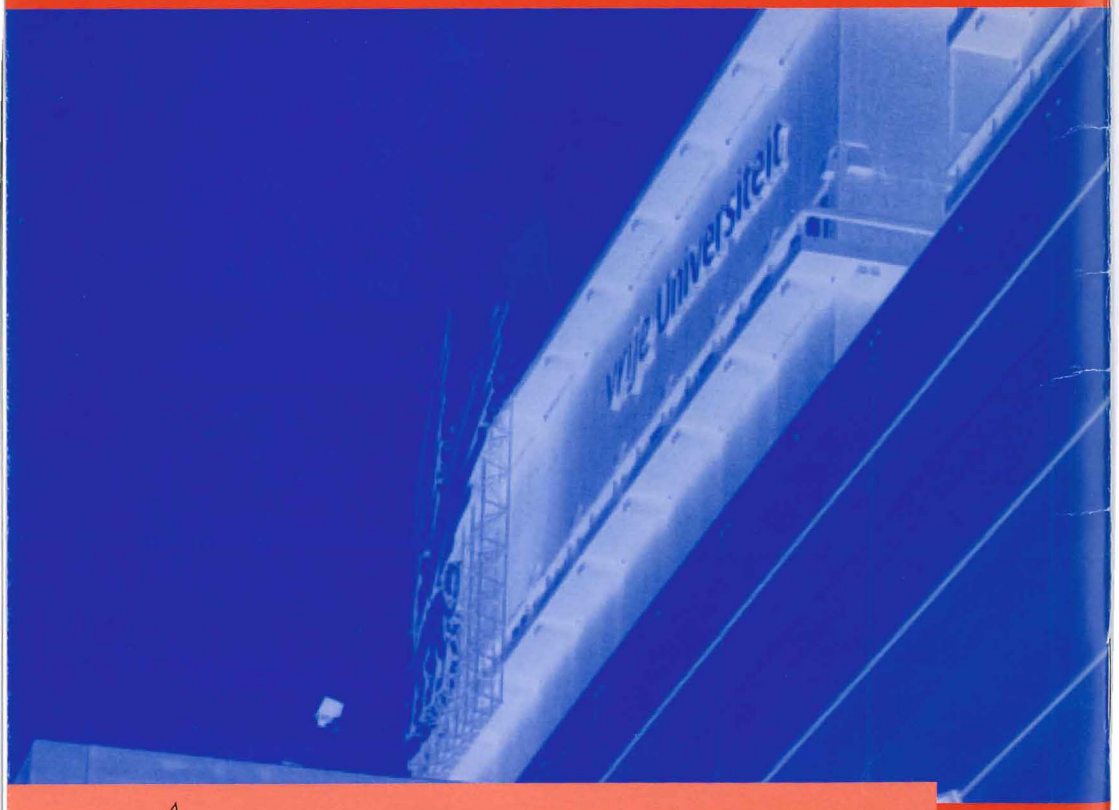
eken doorgeven veronderstellen bevestigen luisteren ve

zoeken doorgeven veronderstellen bevestige

eken doorgeven veronderstellen bevestigen luisteren verwonderen waarnemen verwijzen vergelijk

erzoeken doorgeven veronderstellen bevestigen luisteren

onderen waarnemen verwijzen vergelijken verbinden toetsen onderzoeken doorgeven veronderstell



VU Boekhandel/Uitgeverij Amsterdam
ISBN 90 - 5383 - 853 - 8